



IBM Research

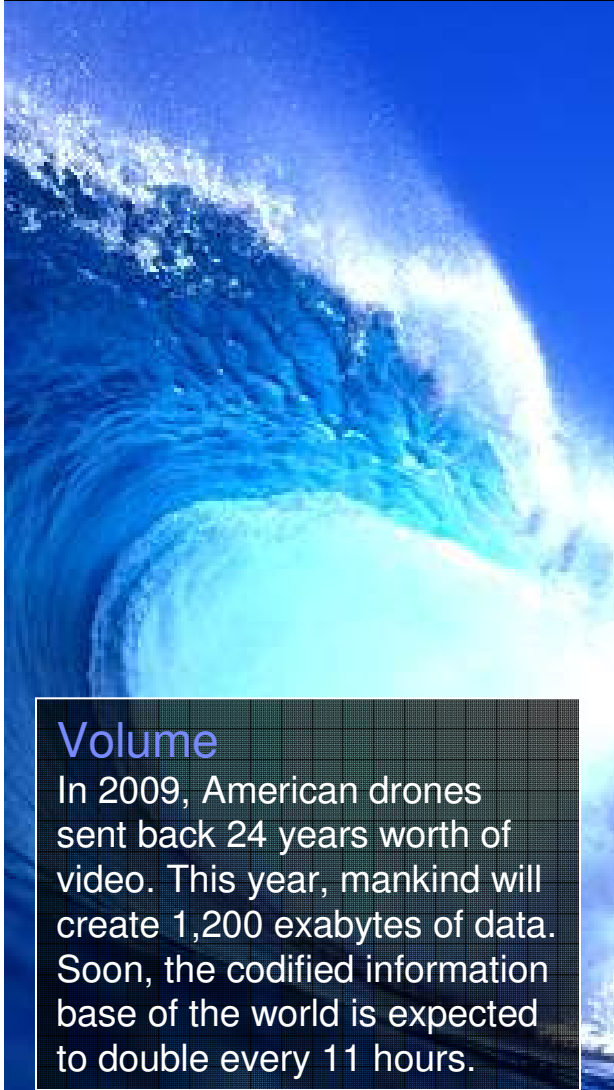
# Stream Computing for Data Fusion and Computational Intelligence

*Chitra Venkatramani*  
*IBM T.J. Watson Research Center*

## Outline

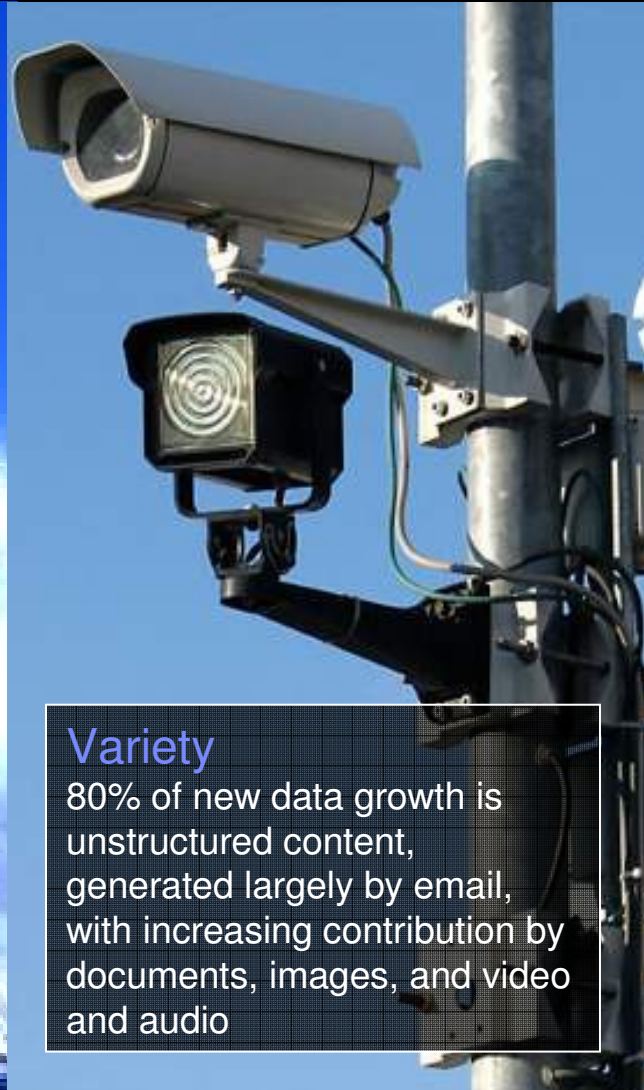
- **The data deluge problem**
- **The stream computing paradigm and Streams**
- **Related Research Activities**

# The world is getting more instrumented and interconnected...



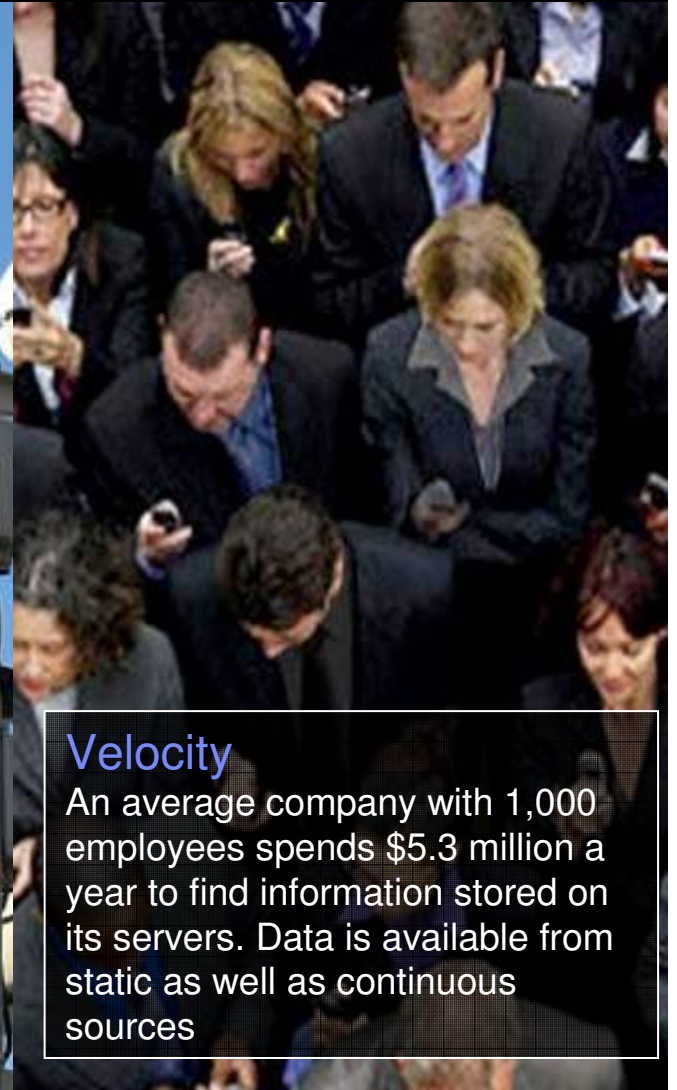
## Volume

In 2009, American drones sent back 24 years worth of video. This year, mankind will create 1,200 exabytes of data. Soon, the codified information base of the world is expected to double every 11 hours.



## Variety

80% of new data growth is unstructured content, generated largely by email, with increasing contribution by documents, images, and video and audio

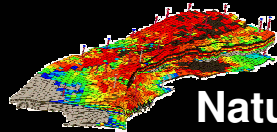


## Velocity

An average company with 1,000 employees spends \$5.3 million a year to find information stored on its servers. Data is available from static as well as continuous sources



# The Need for New Intelligence is Everywhere...



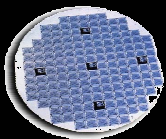
## Natural Systems

- Seismic monitoring
- Wildfire management
- Water management



## Transportation

- Intelligent traffic management



## Manufacturing

- Process control for microchip fabrication



## Health & Life Sciences

- Neonatal ICU monitoring
- Epidemic early warning system
- Remote healthcare monitoring



## Stock market

- Impact of weather on securities prices
- Analyze market data at ultra-low latencies



## Law Enforcement

- Real-time multimodal surveillance



## Fraud prevention

- Detecting multi-party fraud
- Real time fraud prevention

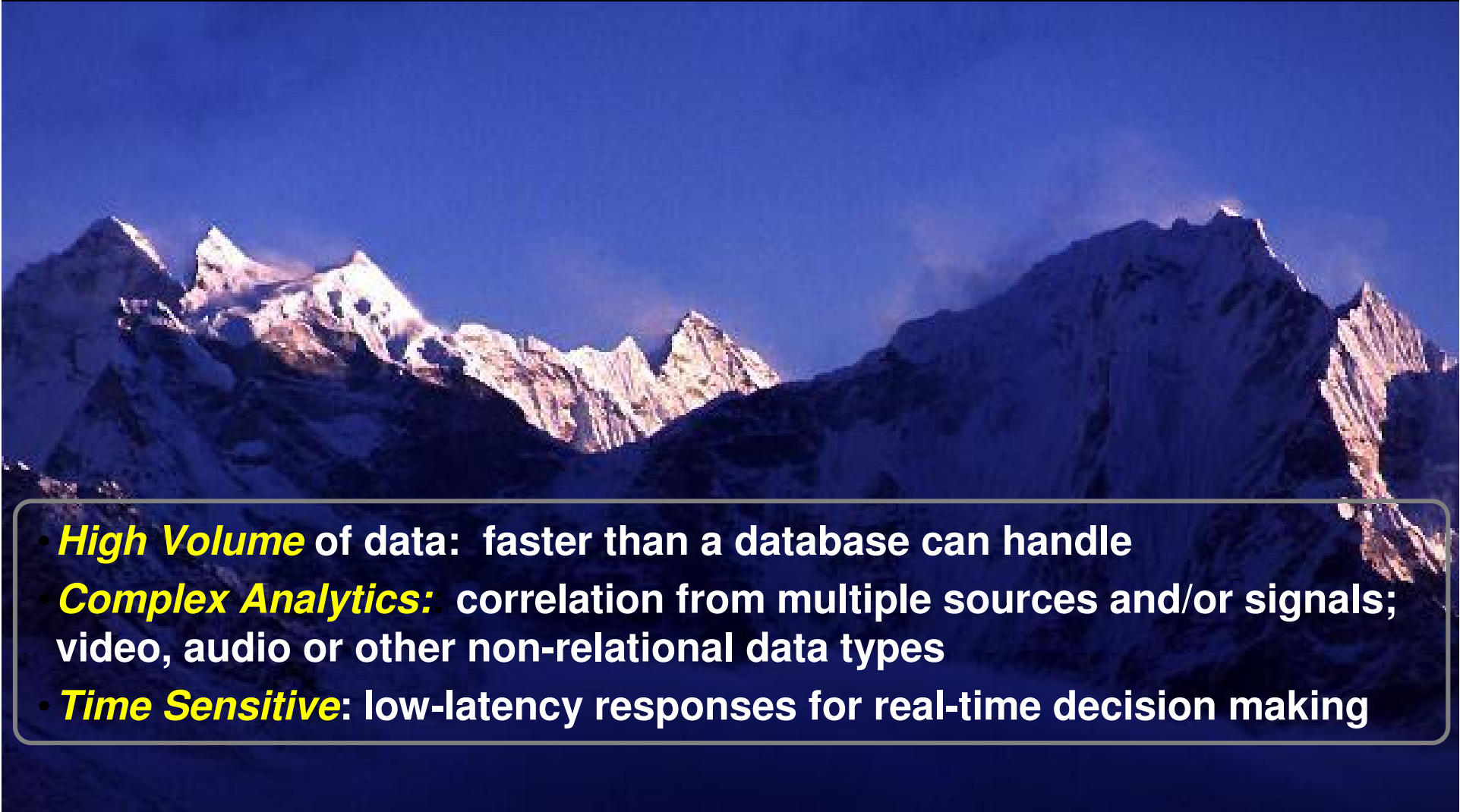


## Radio Astronomy

- Detection of transient events



# The Need for New Intelligence brings New Challenges

- 
- **High Volume** of data: faster than a database can handle
  - **Complex Analytics**: correlation from multiple sources and/or signals; video, audio or other non-relational data types
  - **Time Sensitive**: low-latency responses for real-time decision making

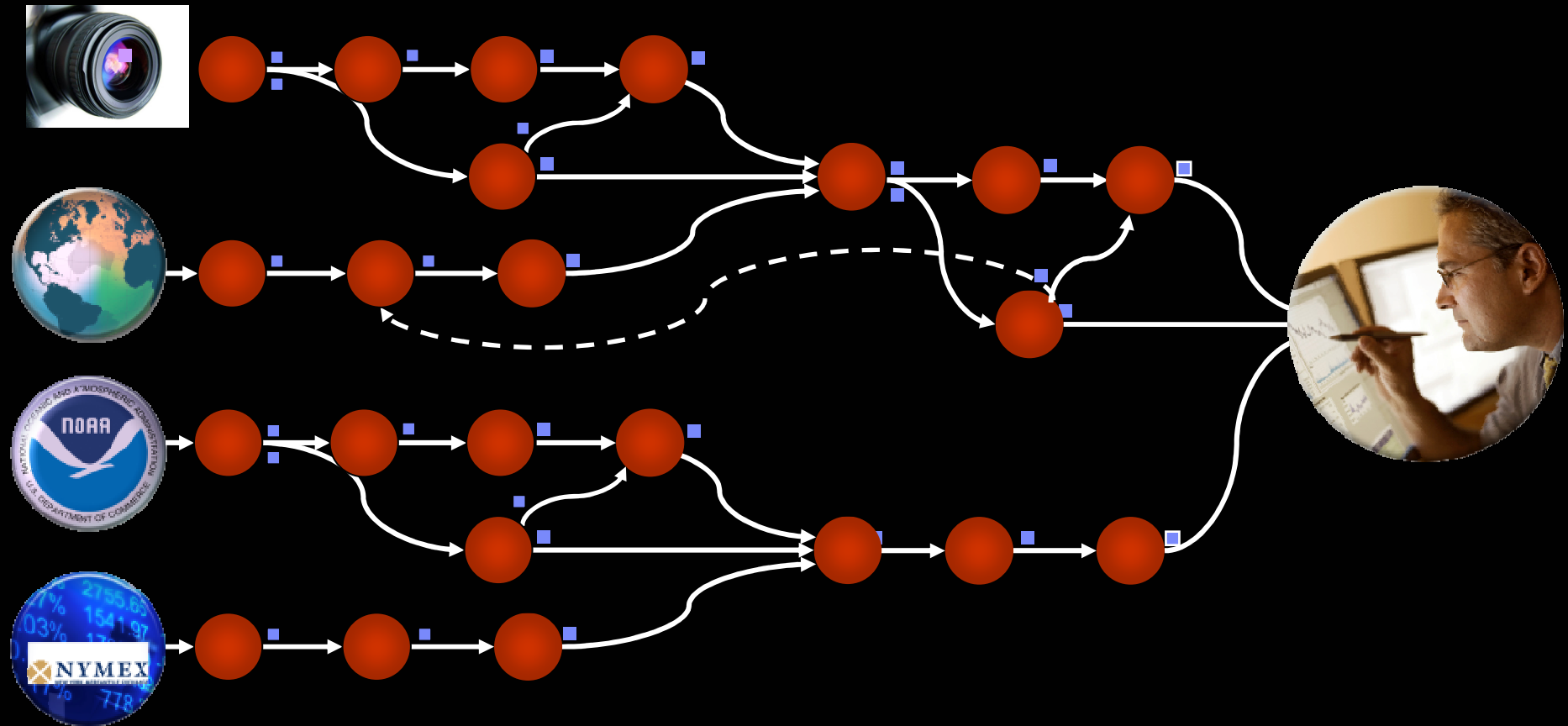
# Application Requirements

- **Streaming Data**
- **Correlation**
  - Connect the dots
  - Reinforce existing knowledge
  - Detect anomalies
- **Hypothesis-driven**
  - Morphable Applications
  - Feedback and control
- **Complex analytics**
  - Structured and unstructured data
  - Resource Adaptive
- **High performance**
  - Throughput, Latency
  - Scalable to keep up with the data-rates
  - Leverage advances in computation and communication

# Stream Computing Illustrated

Continuous Ingestion

Continuous Complex Analysis

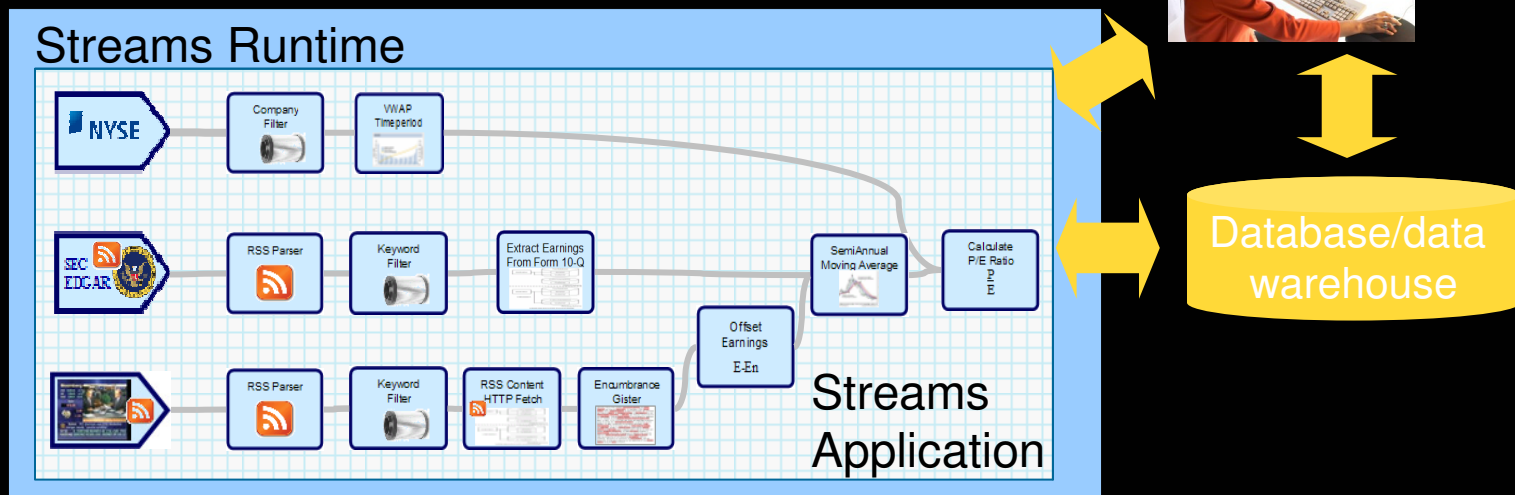


## System S : Commercialized as InfoSphere Streams

- A stream computing software platform to enable better analysis of structured and unstructured data for faster, more informed, and differentiated decision making

*...minimizing time to react*

*Streams applications are composed of analytic operators...*

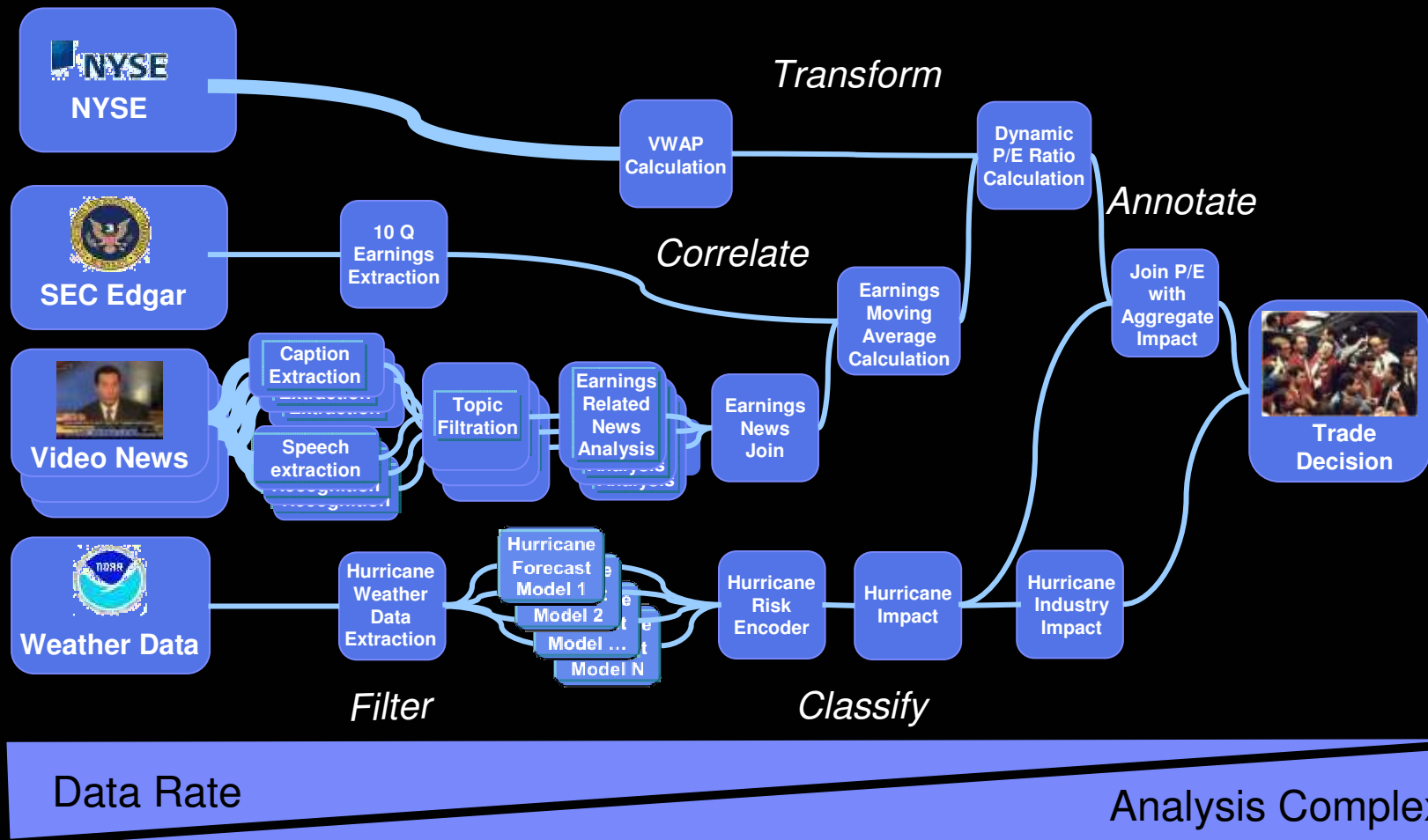


*...processing data as it is continuously generated*

*...extracting and organizing information and intelligence*



# Application Model



# Applications

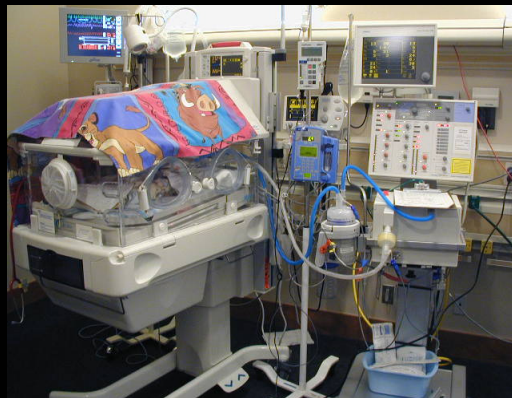
## Surveillance



## Smarter Telecom



## Neonatal Care



## Smart Traffic



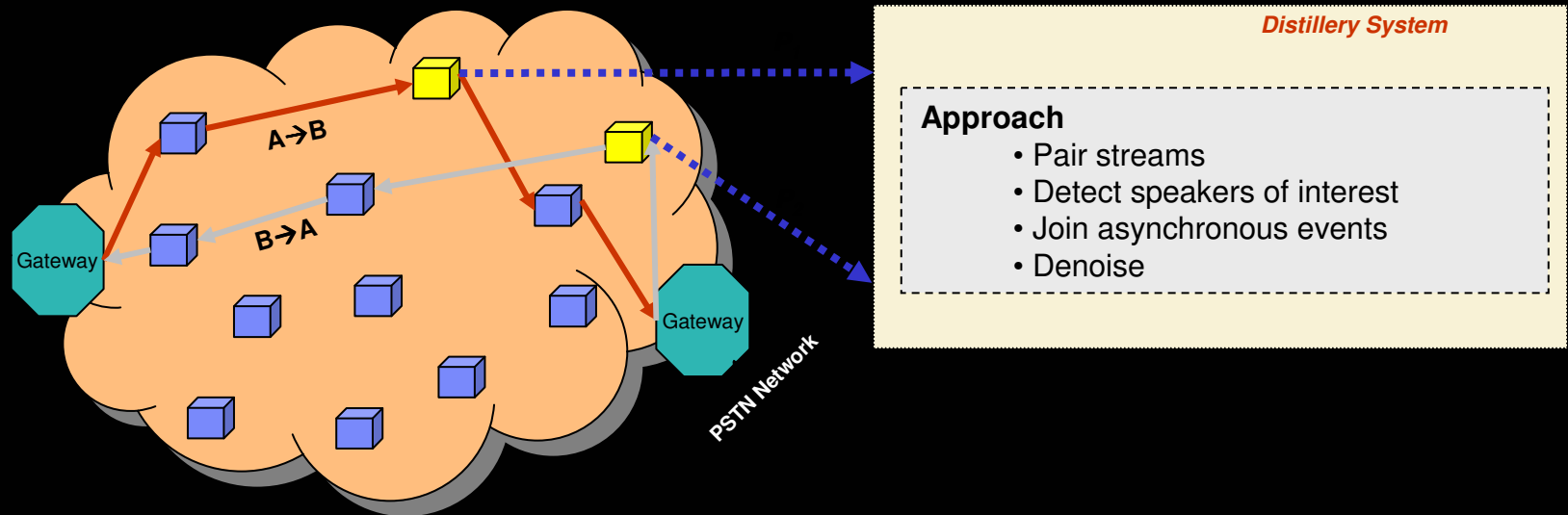
# “Who’s Talking to Whom” Application

## ■ Framework: VoIP Network

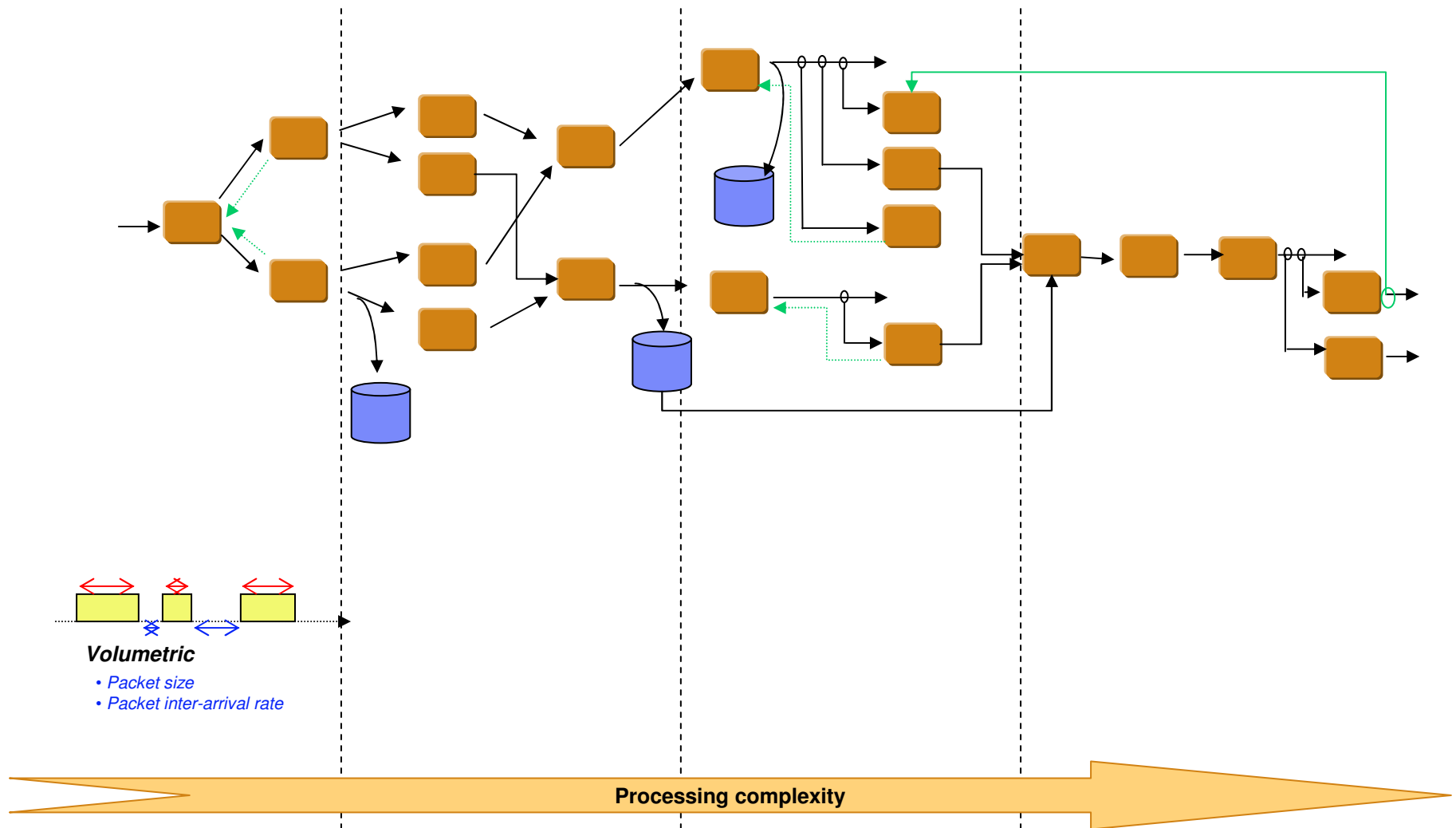
- PSTN Networks and gateways
- K sniffers

## ■ Workload

- 679 speakers
- 2,000+ conversations
- 300+ GSM concurrent streams
  - Tough speech corpus
  - Very low bit rate compression
  - Noisy VAD output
- 15,000+ GSM frames per second

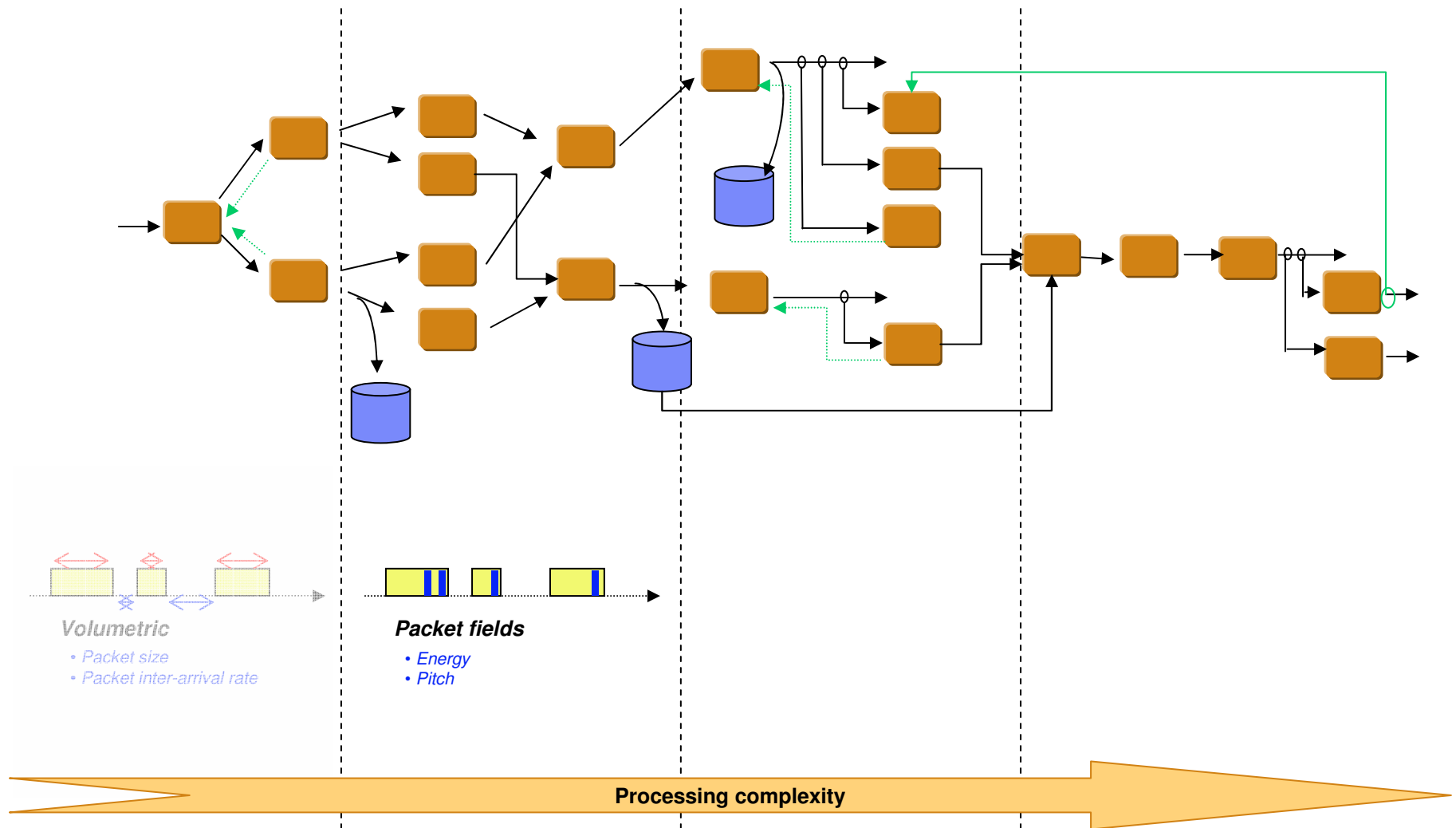


# Incremental Audio Analysis

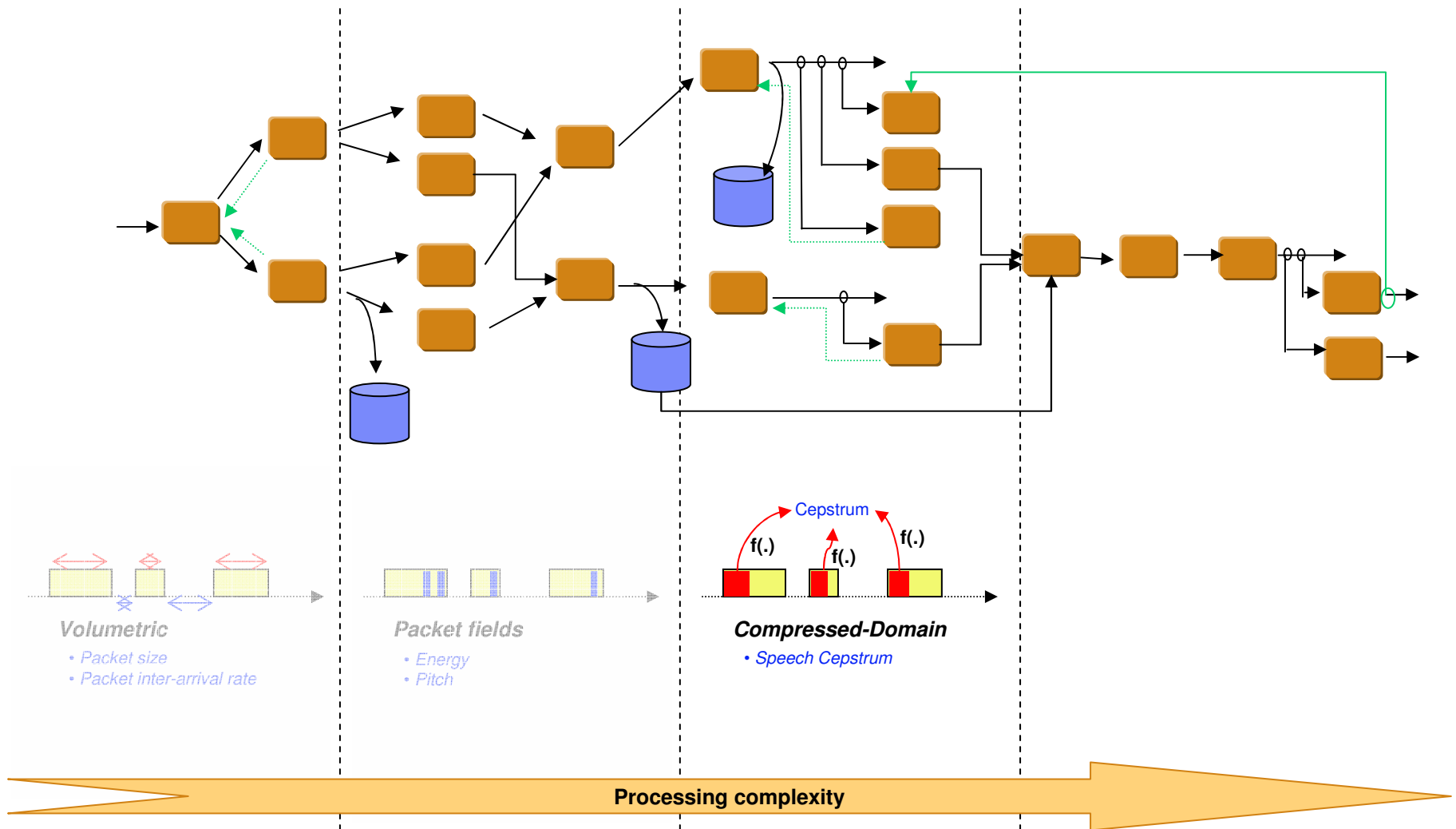




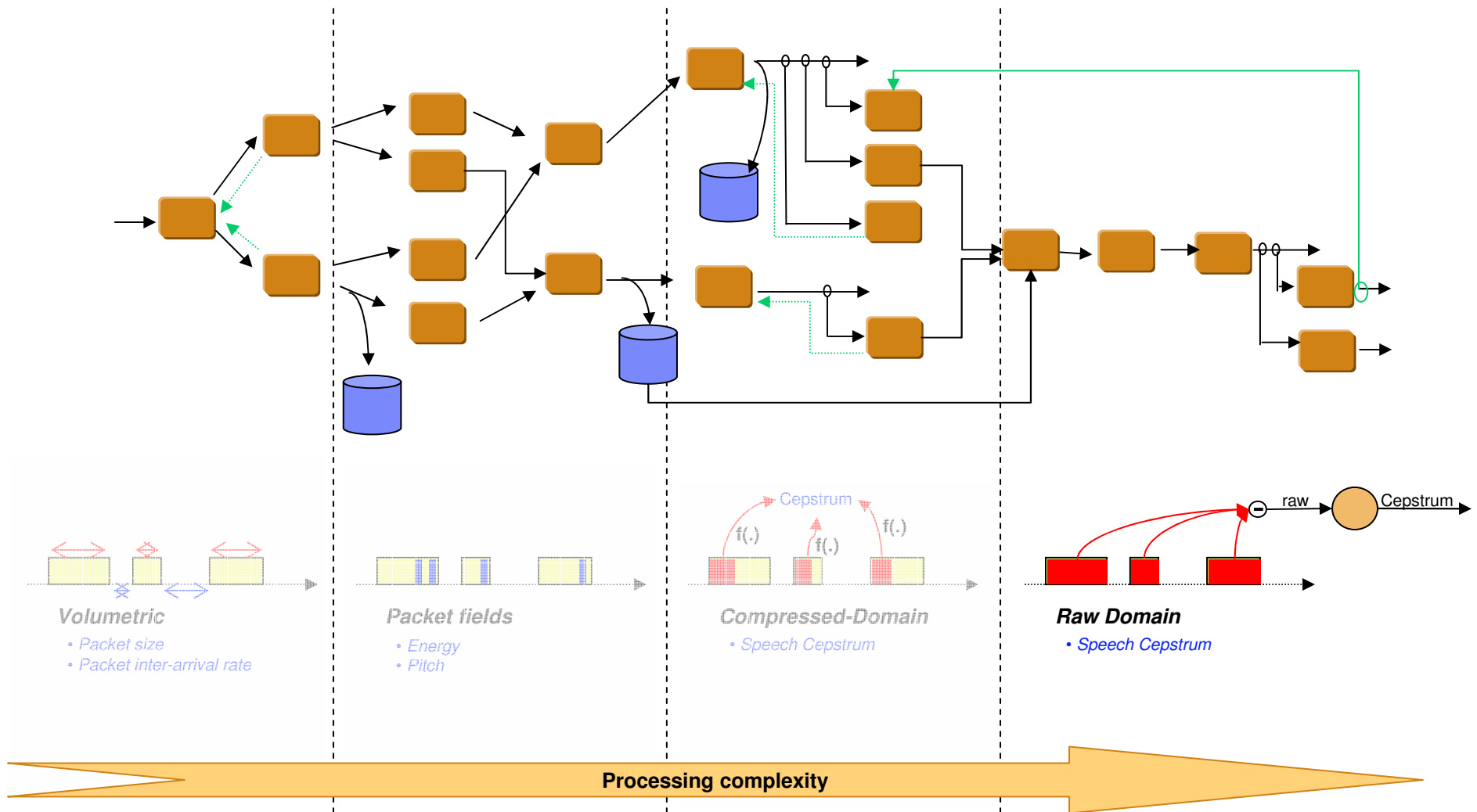
# Incremental Audio Analysis



# Incremental Audio Analysis

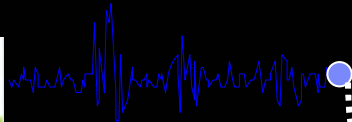


# Incremental Audio Analysis



# Analysis Stages

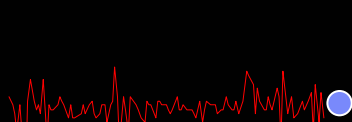
Stream A



Stream B



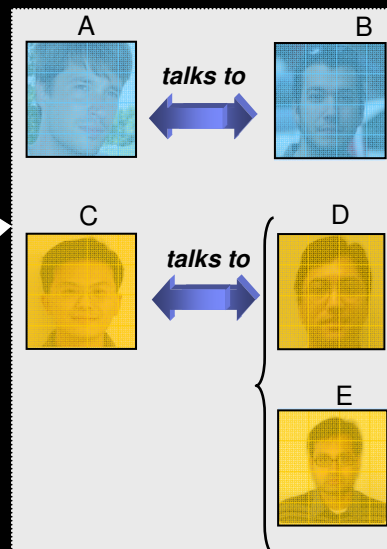
Stream C



Stream D

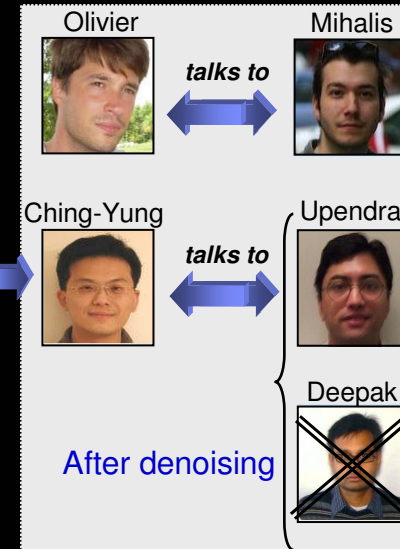


## Conversation Pairing



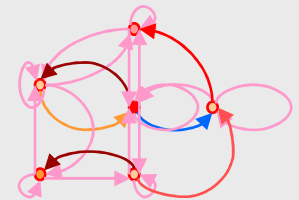
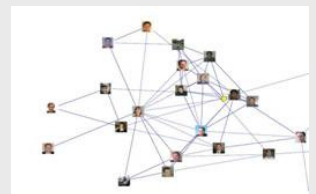
- Just-in-time
- Features: Volumetrics
- Very high accuracy
- Very low complexity
- Robust to noise

## Speaker Detection



- Just-in-time
- Features: GSM domain
- High accuracy
- Moderate complexity
- Robust to noise

## Denoising & Social Network Analysis



- Social network
- Fusion technique
- Iterative method



# Smarter Telco Services

## Customer Requirements

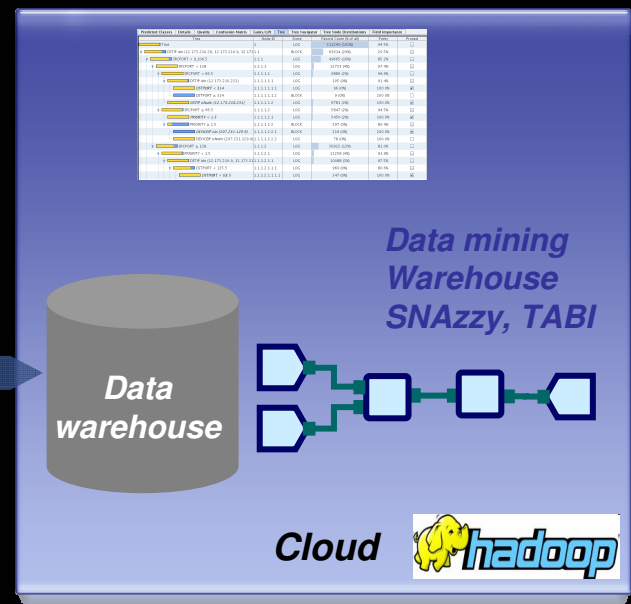
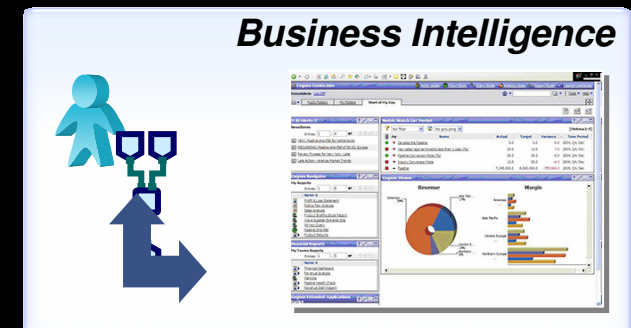
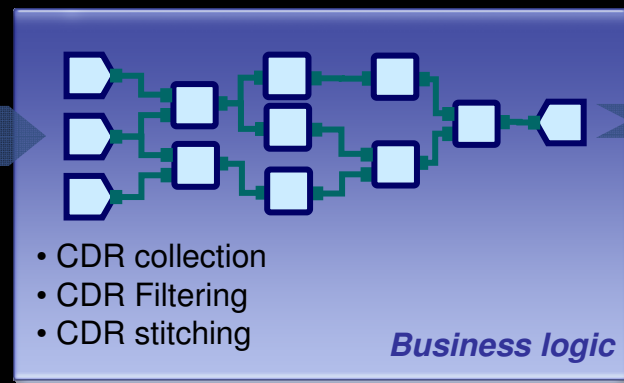
- **Low-latency data processing**
  - Mediation, summarization, monitoring, preprocessing
- **Real-time services**
  - Context based advertizing
  - Real-time campaign management
- **Online data analysis and learning**
  - Churn prediction, model building

## Solution

- **Processing of CDR data using Streams**
  - Extend to other data types
- **Offline data exploration (Warehouse, Hadoop, SNAzzy, TABI)**
  - Tight integration for automated interactions
- **Model scoring and incremental model learning on Streams**

# Current Architecture

Hours to days!



# Stream Processing Architecture

Context

GPS Location, Transaction  
Personal Health Monitor et

Telco Data

Telefon's 7000 Class 5  
Packet SwitchSupport  
data

## Real-time Services

Data Mining  
Scoring EngineData  
summaries

Online Monitoring

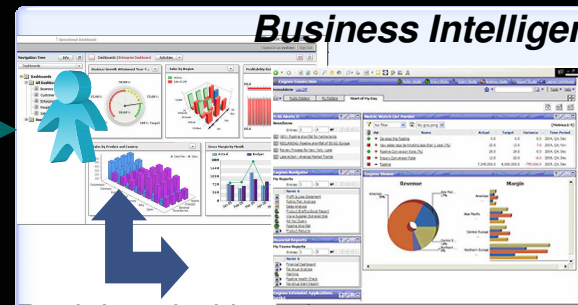
Online Learning with  
Offline Analytics

Mediation

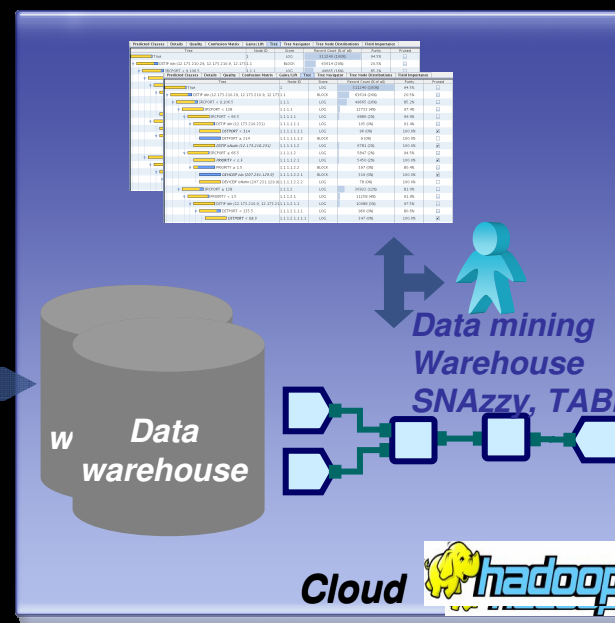
CDR collection  
CDR Filtering  
CDR stitching

Stream Processing

## Business Intelligence



Real-time dashboards



# Efficient Traffic Management

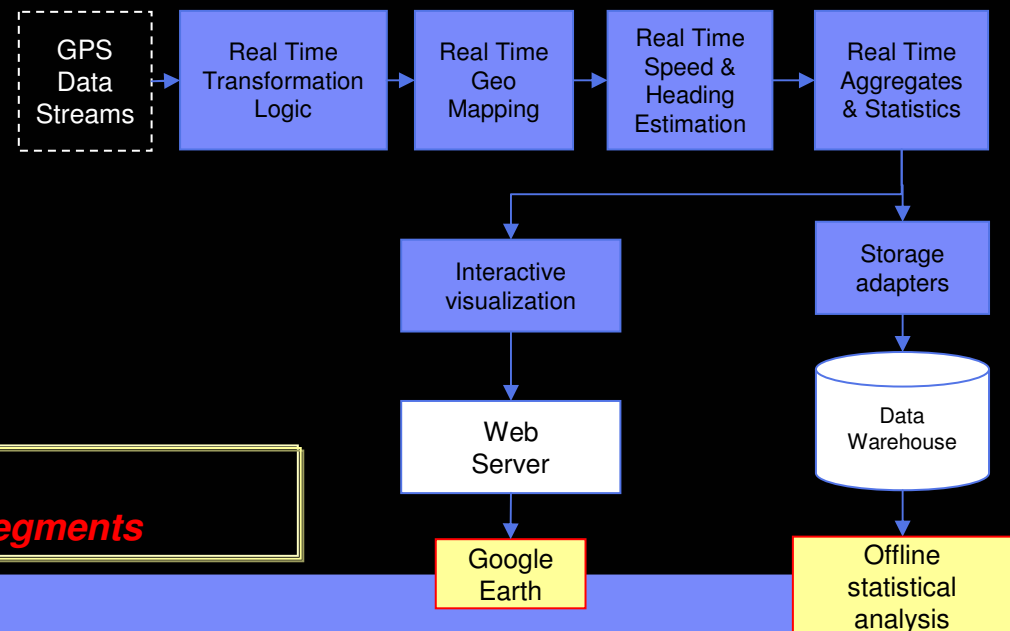


## ■ Multimodal Data Streams

- GPS
- Cell-phones (location tracking),
- Public Transport (bus, docking),
- Pollution measurements,
- Weather Conditions (including road conditions)
- Optical traffic flow detectors,
- Travel time data based on plate recognition,
- Induction loop detector data,
- Accidents in network as they are being recorded,
- Road closures (road work, etc),
- Still pictures from road cameras.



- Real Time Traffic Monitoring
- Real Time Traffic Information
- (Multimodal) Travel Planner



*Only 4 x86 Blade **servers** to process  
250,000 GPS probes **per second**, maps of 630,000 line **segments***

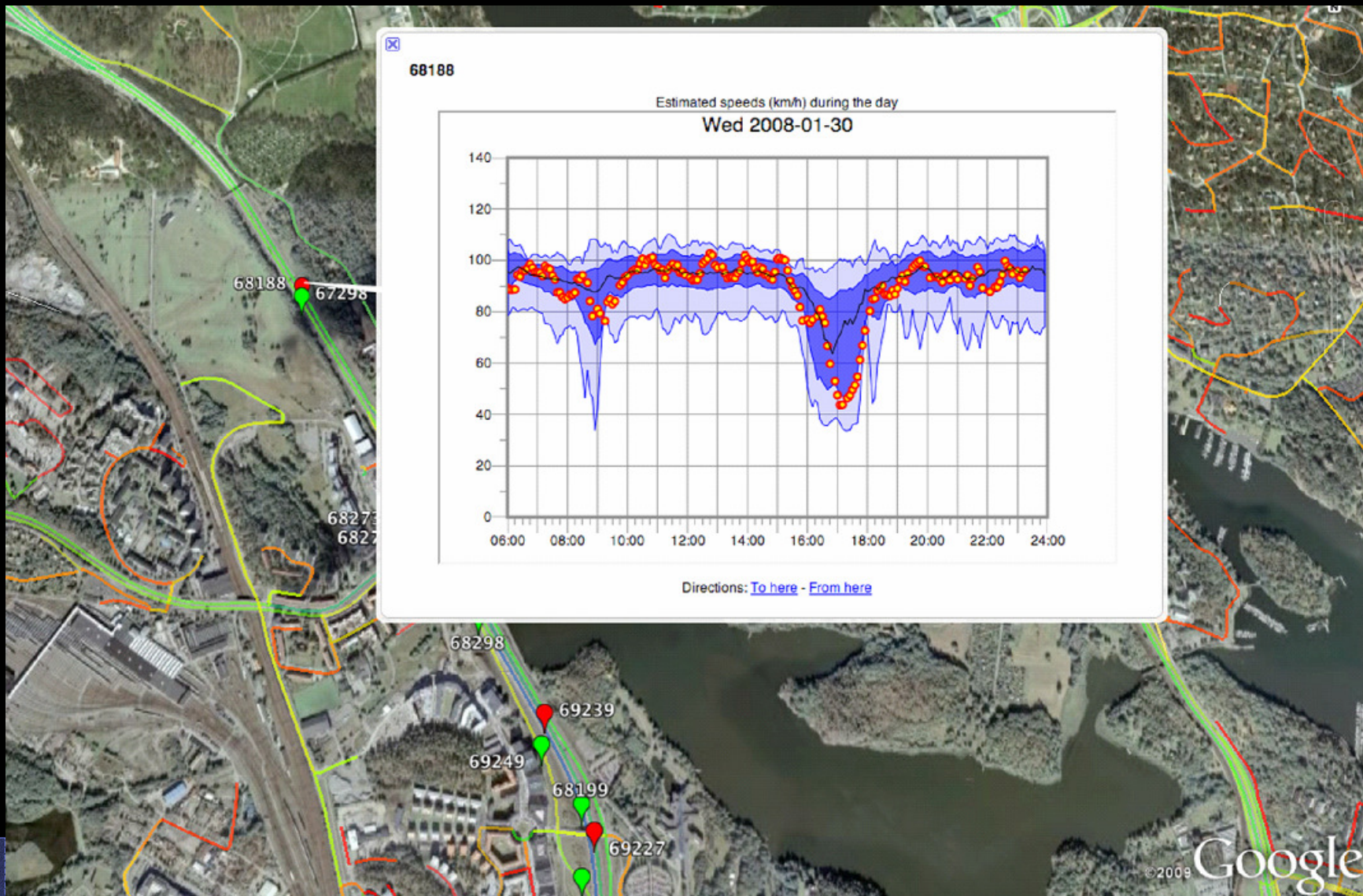


# Real Time Traffic Monitoring

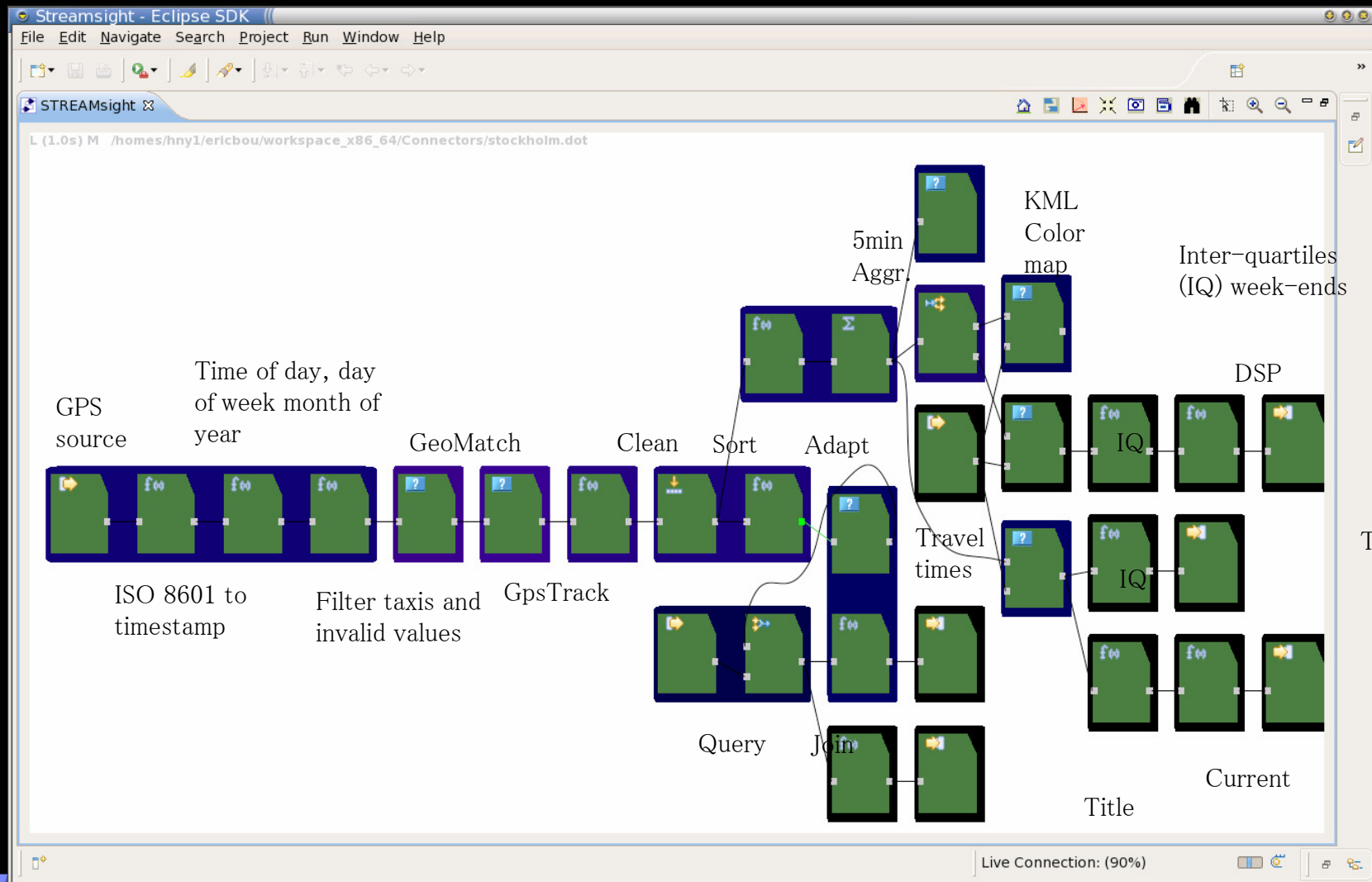




# Real Time Traffic Information



# ITS Application Flow-graph (125k GPS/second)





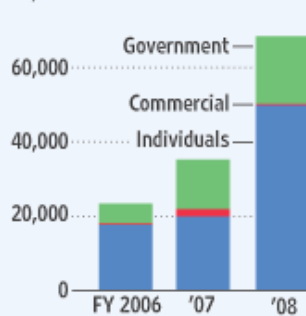
# Cyber-Security – Botnet Detection

## Electricity Grid in U.S. Penetrated By Spies

Wall Street Journal, April 08, 2009

### Stealth Attacks

Number of reported cybersecurity breaches in the U.S., grouped by sector



Note: Fiscal year ends Sept. 30  
Source: Department of Homeland Security

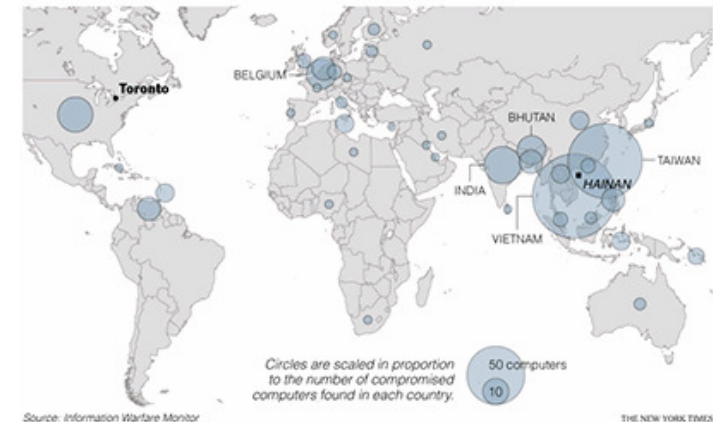
- Cyberspies have penetrated the U.S. electrical grid and **left behind software programs that could be used to disrupt the system**
- The growing reliance of utilities on Internet-based communication has increased the vulnerability of control systems to spies and hackers
- It is nearly **impossible to know who is attacking because of the difficulty in tracking true identities in cyberspace**

## Vast Spy System Loots Computers in 103 Countries

The New York Times, March 28, 2009

### The Vast Reach of 'GhostNet'

Researchers have detected an intelligence gathering operation involving at least 1,295 compromised computers. Below, the locations of 347 of the compromised machines, many of which were tracked to diplomatic and economic government offices of South and Southeast Asian countries.

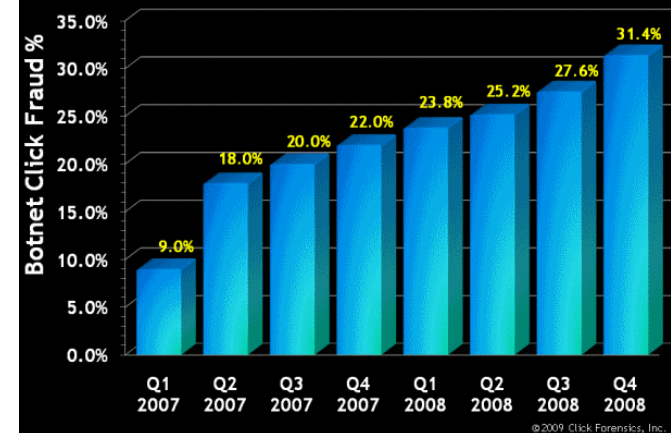


## Biggest Botnets January 2009:

(Wikipedia et al)

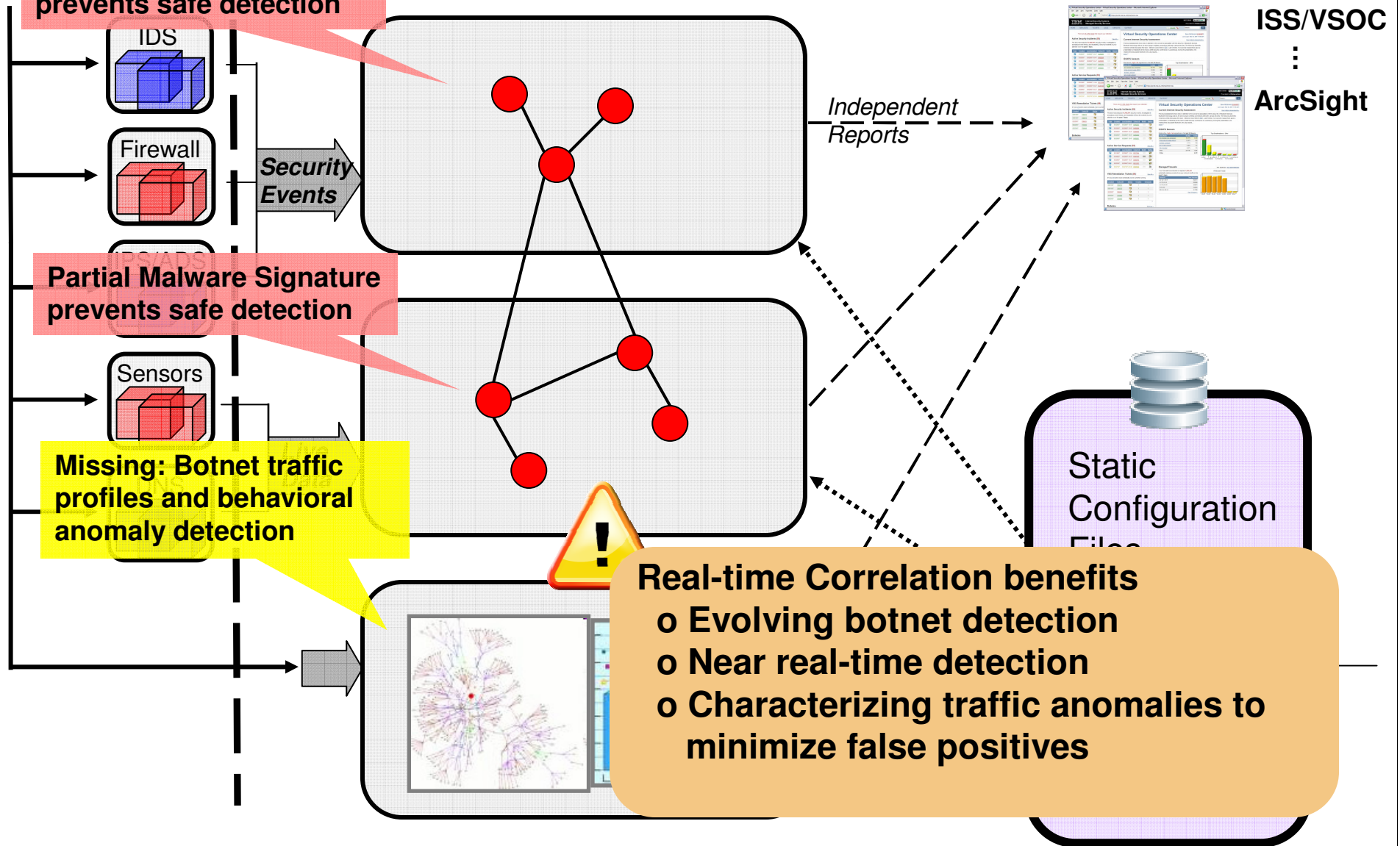
- **Cutwail:** 175,000 infected machines, Type: HTTP Encrypted; Purpose: Spam; Malware: Trojan/Rootkit
- **Rustock:** 130,000 active members per 24 hour period; Type: HTTP Encrypted; Purpose: Spam; Malware: Trojan/Rootkit
- **Donbot:** Size: 125,000 active members per 24 hour period, Type: Custom TCP, Purpose: Spam, Download; Malware: Trojan
- **Ozdok, Xarvester, Grum, Ghag, Cimbot, Waledac, ...**

## Botnet Click Fraud by Quarter

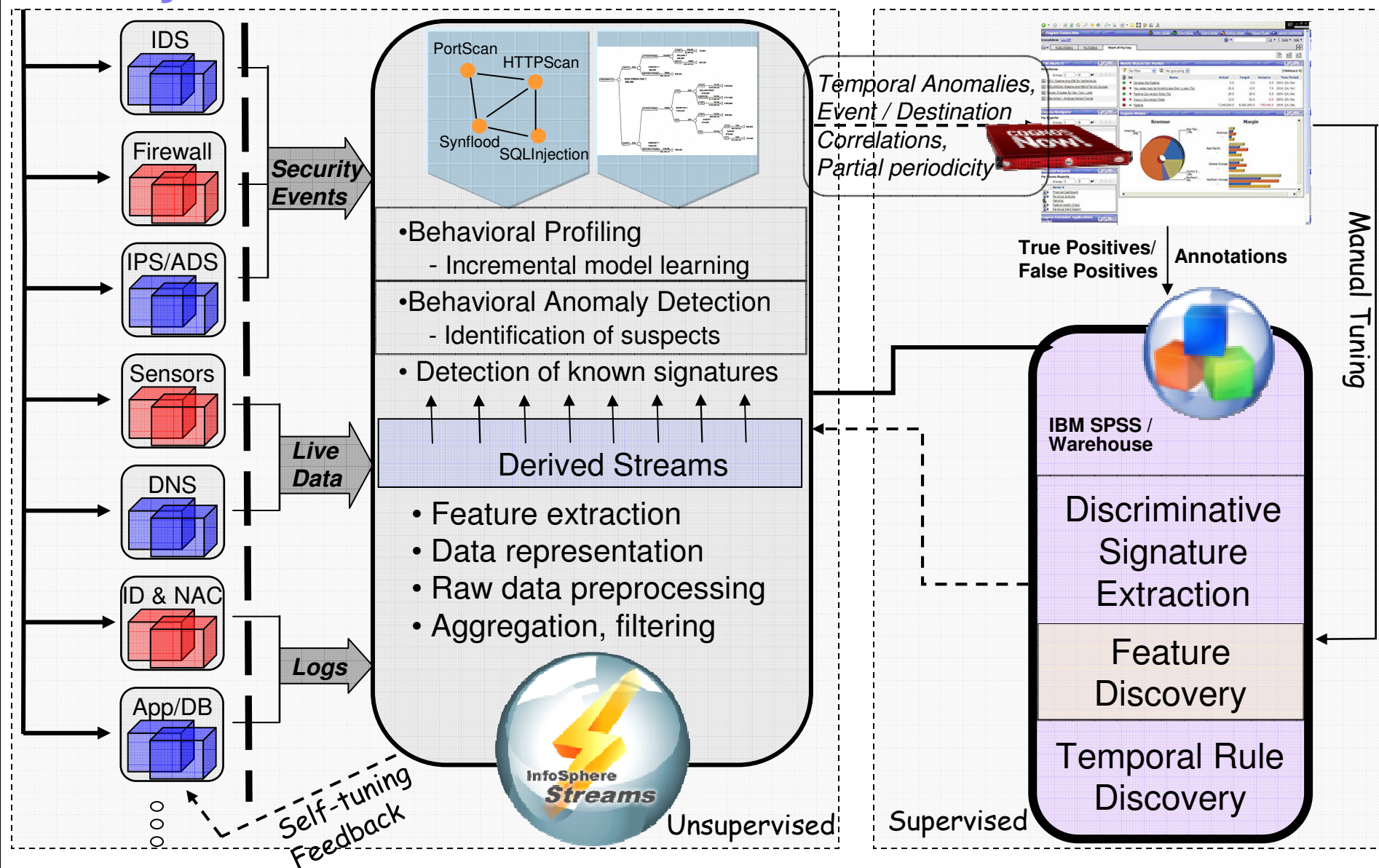




## Limitations



# Analytics Architecture





# Neonatal Care



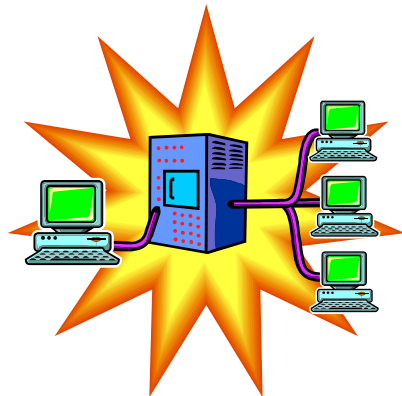
- Multiple devices are attached to the baby or humidicrib
- Medical devices output via serial port in a range of formats
- Indicative readings are recorded on paper every 30 or 60 minutes
- Correlation across multiple sources and episodic conditions make detection of early indicators difficult

<http://preemie.info/cms/modules/news/>

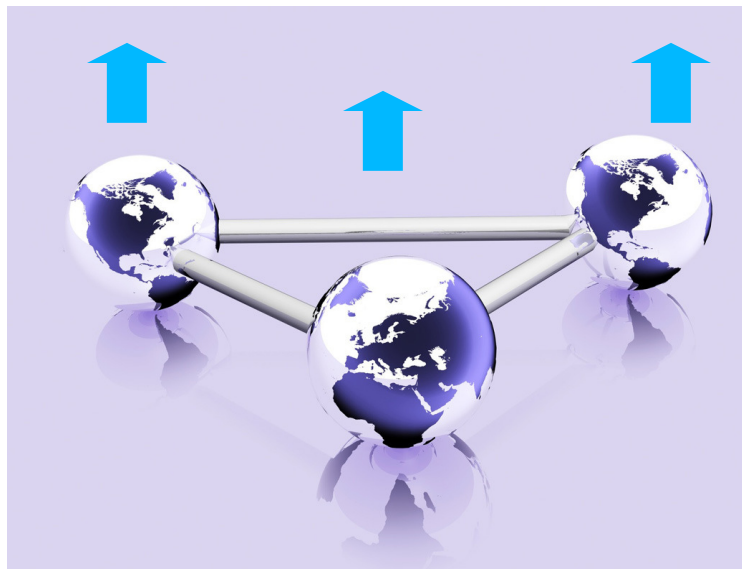
## A 3-Way Collaboration



**IBM Research**



*System and Analytics*



**SickKids**





# The Data Baby Commercial

- Video

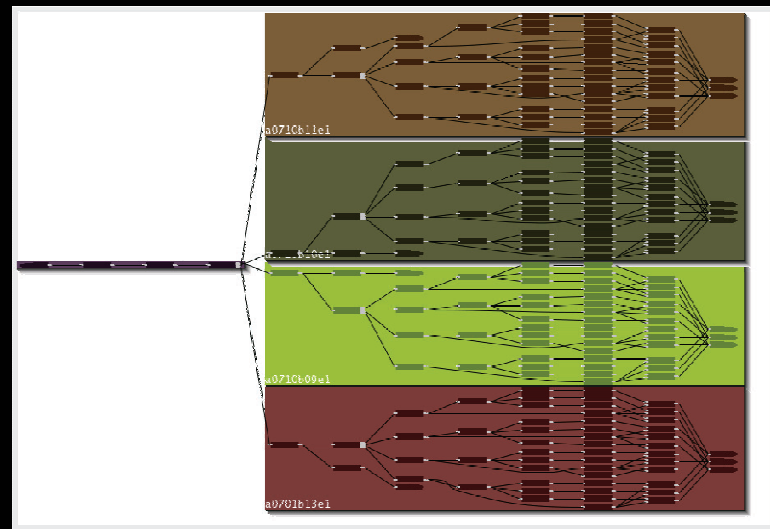


# Language and Compiler

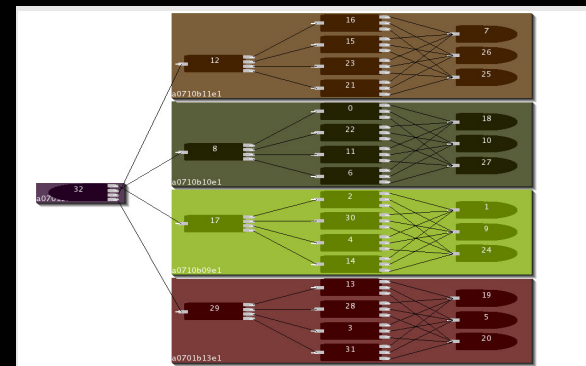
- **Streams Processing Language**
- **Parallelization constructs**
- **Reusable Composite operators**
- **Incremental application composition**
- **Resource hints**
- **Compiler Framework**

# Compiler Framework

- Operator Fusion
  - Fine-grained operators
  - From small parts, make larger ones that fit
- Code generation
  - Generates code to match the underlying runtime environment
    - Number of cores
    - Interconnect characteristics
    - Architecture-specific instructions
  - Driven by automatic profiling
  - Compiler-based optimization
  - Driven by incremental learning of application characteristics



Logical app view



Physical app view

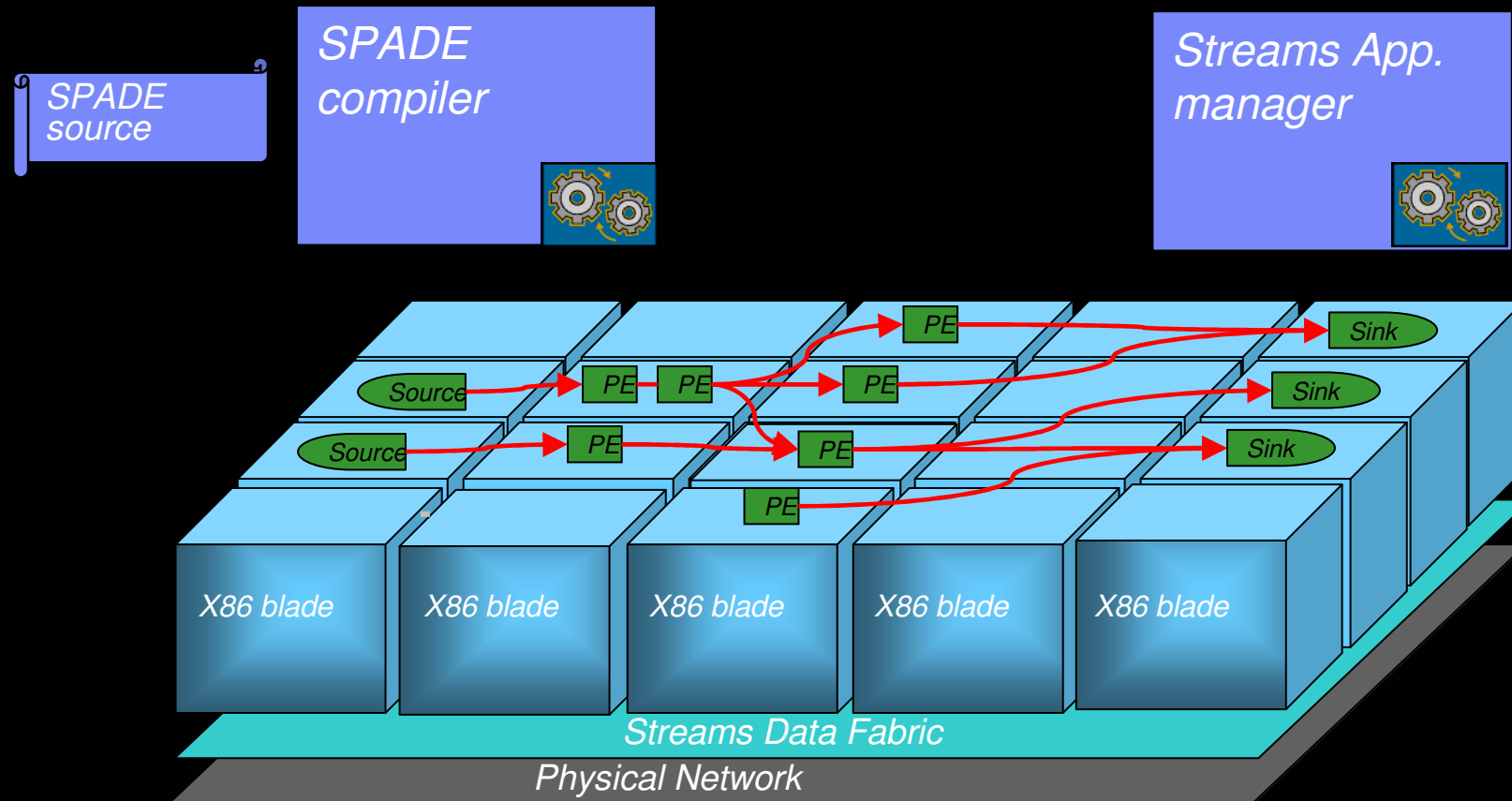


# Runtime

- **Distributed, Scalable**
- **Dynamic Application Composition**
- **Continuous and adaptive resource management**
- **Fault-tolerance**
- **Leverage advances in computation and communication**
  - Multi-core
  - 10GigE, Infiniband



# Runtime



# Tooling

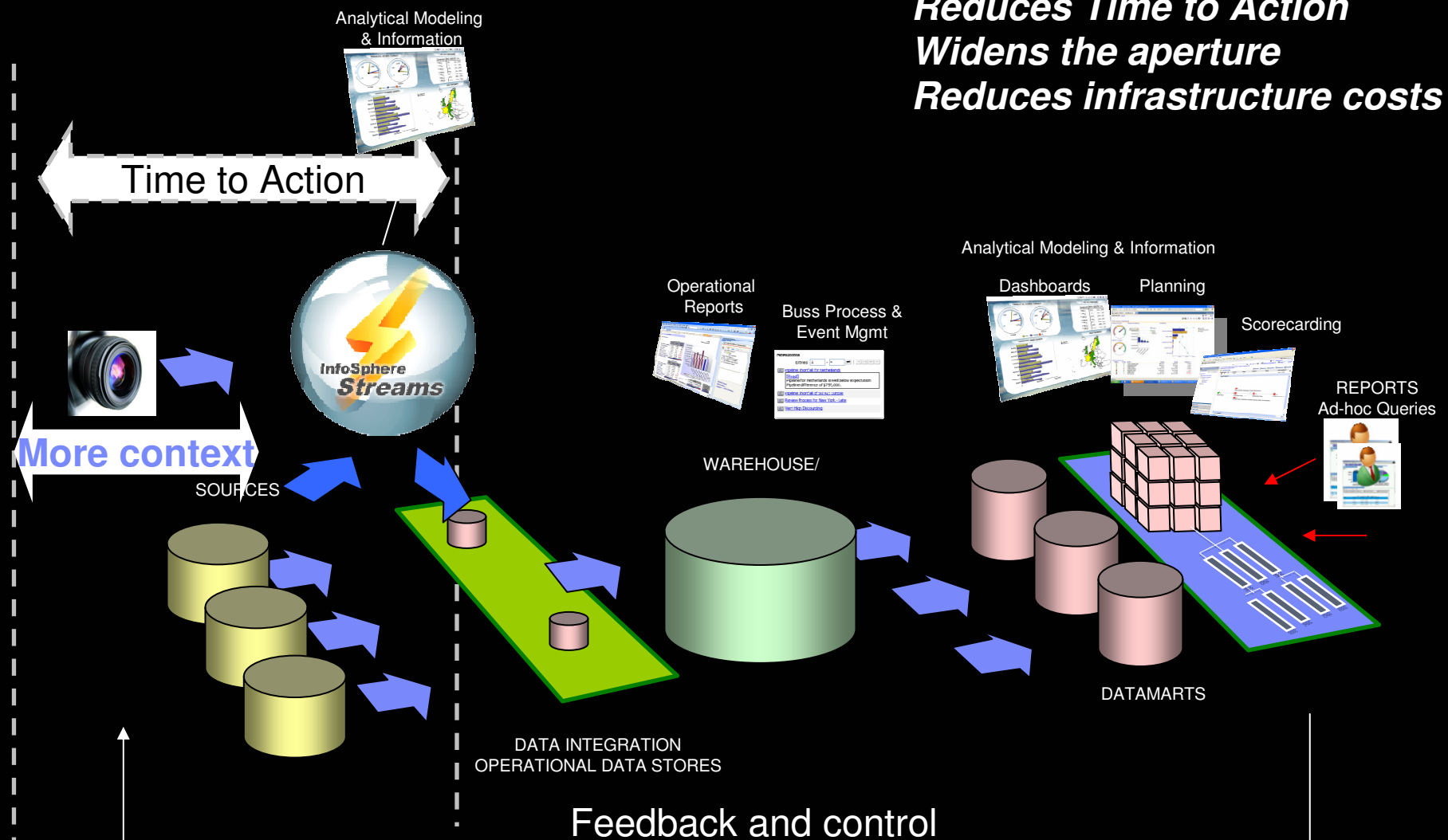
- **Application Development and Debugging**
- **Visualization**
- **Toolkits**
  - Stream-Relational
  - Data mining scoring
  - Time Series
  - Graph Mining
  - Financial
- **Automated Application Composition**

# Analytics

- **Early discard**
- **Incremental Information Exposure**
- **Incremental Analytics**
- **Online learning**
- **Resource-adaptive Analytics**

# Stream computing in the Data Management Eco-system

*Reduces Time to Action*  
*Widens the aperture*  
*Reduces infrastructure costs*



## Current Research

- **Cross-platform Integration**

- High-speed Streaming Ingest into Hadoop/Databases/file-systems
- Workload aware cross-platform Scheduler

- **Dynamic Resource Management**

- Adaptive with dynamic operator fusion
- Streams on cloud, leveraging elastic resource model



# Science of Analytics

- **Addressing the ‘decision overload’ problem**
  - What analyses to perform and when to perform them?
  - What methodologies, algorithms and combination of analytics to use?
  - How to execute the analysis processes?
- **Research to automate much of the analysis process in a principled, scientific way**
  - Knowledge representation and management
  - Models, algorithms for data exploration
  - Analytics Engineering and management
- **Benefits**
  - Analysts focus on strategic, higher-order thinking
  - Developers create new analytic algorithms and tools

# Selected Publications

## R&D Magazine 2010 Top 100 Innovations Award Winner

### Books

- C. Aggarwal. Data Streams: Models and Algorithms, Springer, 2007.

### Applications

- Deepak Turaga, and others. Design Principles for Developing Stream Processing Applications. Software - Practice and Experience Journal, Wiley SP&E. To Appear
- Alain Biem, Eric Bouillet, Hanhua Feng, Anand Ranganathan, Anton Riabov, Olivier Verscheure, Haris N. Koutsopoulos, Carlos Moran: IBM infosphere streams for scalable, real-time, intelligent transportation services. SIGMOD Conference 2010: 1093-1104
- Blount, M.; Ebling, M.R.; Eklund, J.M.; James, A.G.; McGregor, C.; Percival, N.; Smith, K.P.; Sow, D., "Real-Time Analysis for Intensive Care: Development and Deployment of the Artemis Analytic System", IEEE Engineering in Medicine and Biology Magazine, March-April 2010, Vol 29, Issue 2.
- Olivier Verscheure, Michail Vlachos, Aris Anagnostopoulos, Pascal Frossard, Eric Bouillet, Philip S. Yu: Finding "Who Is Talking to Whom" in VoIP Networks via Progressive Stream Clustering. ICDM 2006: 667-677
- H. Tseng, O. Verscheure, D. S. Turaga and U. Chaudhari, "Optimal quantization for adapted GMM-based speaker verification," IEEE Interspeech 2007.
- D. S. Turaga, O. Verscheure, J. Wong, L. Amini, G. Yocum, E. Begle, B. Pfeifer, "Online FDC Control Limit Tuning with Yield Prediction using Incremental Decision Tree Learning," AEC/APC Symposium, 2007
- Ching-Yung Lin, Olivier Verscheure, Lisa Amini: Semantic Routing and Filtering for Large-Scale Video Streams Monitoring. ICME 2005: 1408-1411
- Deepak S. Turaga, Brian Foo, Olivier Verscheure, Rong Yan: Configuring topologies of distributed semantic concept classifiers for continuous multimedia stream processing. ACM Multimedia 2008: 289-298
- F. Fu, D. S. Turaga, O. Verscheure, M. Van der Schaar and L. Amini, "Configuring Competing Classifier Chains in Distributed Stream Mining Systems," IEEE J. Selected Topics in Signal Processing.

# Publications

## Language, Runtime

- Robert Soulé, and others A Universal Calculus for Stream Processing Languages, European Symposium on Programming, ESOP, 2010.
- Rohit Khandekar and others. COLA: Optimizing Stream Processing Applications Via Graph Partitioning. ACM/IFIP/USENIX International Middleware Conference, Middleware, 2009.
- Gabriela Jacques-Silva and others. Language Level Checkpointing Support for Stream Processing Applications. International Conference on Dependable Systems and Networks. IEEE/IFIP DSN 2009.
- Xiaolan J. Zhang, and others. Implementing a High-Volume, Low-Latency Market Data Processing System on Commodity Hardware using IBM Middleware. Workshop on High Performance Computational Finance at SC09, Nov, 2009.
- Buğra Gedik, Henrique Andrade, and Kun-Lung Wu. A Code Generation Approach to Optimizing High-Performance Distributed Data Stream Processing. International Conference on Information and Knowledge Management, ACM CIKM, 2009.
- Scott Schneider, Henrique Andrade, Buğra Gedik, Alian Biem, and Kun-Lung Wu. Elastic Scaling of Data Parallel Operators in Stream Processing. International Parallel and Distributed Processing Symposium. IEEE IPDPS 2009.
- Shicong Meng, Srinivas R. Kashyap, Chitra Venkatramani, Ling Liu: REMO: Resource-Aware Application State Monitoring for Large-Scale Distributed Systems. ICDCS 2009
- N. Bansal, R. Bhagwan, N. Jain, Y. Park, D. S. Turaga, C. Venkatramani, "*Towards Optimal Operator Placement in Partial-Fault Tolerant Applications*", IEEE Infocom 2008, April, Phoenix, AZ

## Tooling

- Wim De Pauw, Henrique Andrade, Lisa Amini: Streamsight: a visualization tool for large-scale streaming applications. SOFTVIS 2008: 125-134
- Buğra Gedik, Henrique Andrade, Andy Frenkiel, Wim De Pauw, Michael Pfeifer, Paul Allen, Norman Cohen, and Kun-Lung Wu. Debugging Tools and Strategies for Distributed Stream Processing Applications. Software - Practice and Experience Journal, Wiley SP&E. Volume 39 Issue 16, 2009.
- Eric Bouillet, Mark Febowitz, Zhen Liu, Anand Ranganathan, Anton Riabov: A tag-based approach for the design and composition of information processing applications. OOPSLA 2008: 585-602

# Thank You!